Robust Multi-View Change Detection*

Alessandro Lanza Luigi Di Stefano
University of Bologna – DEIS – ARCES
40136 Bologna, Italy
{alanza,ldistefano}@deis.unibo.it

Jérôme Berclaz François Fleuret Pascal Fua
EPFL – CVLAB
CH-1015 Lausanne, Switzerland
{jerome.berclaz,francois.fleuret,pascal.fua}@epfl.ch

Abstract

We present a multi-view change detection approach aimed at being robust with respect to common "disturbance factors" yielding image changes in real-world applications. Disturbance factors causing "slow" or "fast-and-global" image variations, such as light changes and dynamic adjustments of camera parameters (e.g. auto-exposure and auto-gain control), are dealt with by a proper single-view change detector run independently on each view. The computed change masks are then fused into a "synergy mask" defined into a common virtual top-view, so as to detect and filter-out "fast-and-local" image changes due to physical points lying on the ground surface (e.g. shadows cast by moving objects and light spots hitting the ground surface).

1 Introduction

Detecting changes in video sequences plays a crucial role in many computer vision applications since the performance of higher-level processing modules, such as objects tracking and classification, often relies on the accuracy of the computed change masks. In the space of all the possible image changes a *good* change detector should be able to discriminate between "semantic" (i.e. due to variations of the scene geometry) and "appearance" (i.e. due to other causes, that we call "disturbance factors") changes. In particular, a change detection algorithm should be robust with respect to disturbance factors arising both in the imaged scene (e.g illumination changes) and in the imaging device (e.g. noise, dynamic adjustments of device parameters such as auto-exposure and auto-gain control).

Most of the single-view change detectors proposed in literature (e.g. [3], [10]) can deal effectively with camera noise and "slow" scene appearance changes (e.g. scene illumination changes due to time of the day). To this purpose, a temporally adaptive per-pixel statistical modelling of the scene background appearance is exploited. To avoid the inclusion of foreground objects in the background appearance model, the model adaptation rate must be chosen accurately, depending on the foreground objects foreseen velocity. In particular, the lower the foreground objects foreseen velocity, the lower the background model adaptation rate. Hence, in general only quite slow appearance changes can be dealt with by these algorithms. Some approaches have been proposed (e.g. [2],[7],[9],[11])

*This work was supported by the Indo Swiss Joint Research Programme (ISJRP).

which can deal effectively also with "fast-and-global" scene appearance changes, that is fast changes modifying pixel intensities by a unique mapping function. Examples of such changes are those due to fast-and-global scene illumination changes (e.g. light switches, a cloud passing by the sun) and to dynamic adjustments of camera parameters (e.g. auto-exposure and auto-gain control). "Fast-and-local" scene appearance changes (e.g. shadows cast by moving objects, light spots hitting a nearly lambertian surface) are a hard-to-solve problem for single-view approaches.

Multi-view change detection can exploit more information and therefore deal more effectively with disturbance factors. As regards the way information is exploited, we define:

c.1) *temporal consistency* constraint: given a view-point *v*, the processed frames are images of the same scene taken at different times;

c.2) *spatial coherence* constraint: given a time *t*, the processed frames are images of the same scene taken from different view-points;

By applying only the spatial coherence constraint the basic multi-view change detection approach is carried out. In practice, at each time *t* the output is computed by comparing all the simultaneous images captured from the different view-points. However, all the available information can be exploited by applying both the constraints. This is in theory the most effective approach. We present a multi-view change detection algorithm of this type. In particular, we apply the temporal consistency constraint as a first processing step by carrying out single-view change detection on each original view. Then, the spatial coherence constraint is applied by "fusing" the single-view change masks into a virtual top-view. Such an approach allows for filtering-out the appearance changes due to the major disturbance factors, including sudden-and-local illumination changes.

The paper is organized as follows. In section 2 the state-of-the-art in multi-view change detection is outlined. The proposed algorithm is presented in section 3. Experimental results are discussed in section 4, conclusions are drawn in section 5.

2 Related Work

In [5] a "lighting independent" multi-view change detection algorithm is presented. Stationarity of the capturing devices as well as of the scene background surface geometry is assumed, so that the geometric transformations warping one of the views, called "primary" view, into all the other "auxiliary" views can be computed off-line. On-line, just the change mask in the primary view is computed. Moreover, only the spatial coherence constraint is applied. In practice, at each time, the colour of every pixel in the primary view is compared with the colour of corresponding pixels in the auxiliary views, using the geometric transformations. If colour is similar, according to a simple metric consisting in the absolute value of the Euclidean distance, the pixel in the primary view is marked as background; otherwise, it is marked as foreground. This approach inherently suffers from both false and missed detections. False detections, called "occlusion shadows", occur when a background pixel in the primary view is occluded by a foreground object in the auxiliary view. Missed detections occur when an evenly coloured foreground object occludes a pair of corresponding pixels, for colour being very similar. The authors propose to filter-out false detections by using more than two views (at least two auxiliary

views) and ANDing the binary masks attained by comparing the primary view to each of the auxiliary views. However, they do not discuss how to deal with missed detections.

The work in [8] is aimed at improving the approach proposed in [5]. As in [5], the change mask in the primary view is computed by applying only the spatial coherence constraint. However, the following improvements are proposed:

a) a slightly more complex and effective metric (i.e. a normalized colour difference averaged on a $n \times n$ neighborhood of pixels) is used to measure colour similarity between corresponding pixels in different views;

b) the false detections problem is addressed from a sensor planning perspective. In particular, it is shown how occlusion shadows can be removed by using just two views, provided that a suitable configuration of the capturing devices is adopted;

c) the missed detections problem is tackled as well. The particular sensors configuration adopted to filter-out occlusion shadows yields missed detections localized only at the lower portion of each detected foreground blob. This is exploited to fill-in possible missed detections by means of a quite complex heuristic procedure.

Both [5] and [8] rely on the assumption that a patch of the scene background surface yields a very similar colour into simultaneous images taken from different view-points. If this is true, invariance to temporal changes of the radiance emitted by the scene background surface (i.e. to slow or fast and global or local scene illumination changes) is achieved, since such changes will affect simultaneous views identically. However, in practice this assumption may not be satisfied. In fact, dynamic adjustments of the camera parameters (e.g. auto-gain and auto-exposure control) may occur in the different views at different times and by a different intensity mapping function. These adjustments cannot be handled inherently by either [5] or [8]. In turn, [5] recommends explicitly to disable the auto-gain mechanism of the capturing devices. However, disabling these dynamic adjustment mechanisms is a strong limitation in many practical applications, especially as regards outdoor installations.

The most related work to our approach is presented in [6]. It is focused on tracking but relies on multi-view change detection as the first processing step. People moving on a ground plane are tracked by their ground locations, that is feet. At each processing time feet are detected by a multi-view change detection approach, that we call here "change maps fusion": the ground plane homographies warping a reference view into each of the other views are inferred off-line. On-line, single-view change detection is carried out independently on each view to compute a change probability map. To this purpose, a well-known background subtraction algorithm based on statistical temporally adaptive background modelling by mixture of gaussians is deployed ([10]). Hence, the computed change probability maps are warped in the reference view by using the inferred homographies and then multiplied together, thus attaining a "synergy map". It is easy to understand how this map gives, for each pixel in the reference view, the probability to be the image of a ground plane patch for which the emitted radiance is changed (with respect to the current appearance background model and according to the chosen single-view change detection algorithm). Finally, the synergy map is thresholded. By this procedure, the authors assume to detect only the ground plane locations of people, that is their feet. Hence, feet are tracked in the reference view by a spatio-temporal clustering approach (graph cuts). However, the proposed use of the change maps fusion approach

will inherently detect as foreground not just feet but also other possible ground plane appearance changes, such as shadows cast by moving objects on the ground plane or light spots hitting the ground plane. In fact, such changes are not filtered-out by the single-view change detection approach in [10].

3 The proposed algorithm

We assume stationarity of the capturing devices as well as of the scene background surface geometry, so that geometric registration of background over different views can be computed off-line. Moreover, we take into consideration a planar background, hereinafter called "ground plane". Hence, for each original view v, we infer off-line the homography $H^v : \mathbb{R}^2 \ni \boldsymbol{p}^v \mapsto \boldsymbol{p}^T \in \mathbb{R}^2$ warping each pixel $\boldsymbol{p}^v$ imaging a ground plane patch in the original view into the pixel $\boldsymbol{p}^T$ imaging the same patch in a common virtual top-view T. By considering a set of $N > 4$ original view $\leftrightarrow$ top-view points correspondences, the homographies are inferred by least squares regression. A data normalization procedure is adopted to make the necessary matrix calculations less prone to numerical errors ([4]).

As far as on-line processing is concerned (Figure 1), at each time t first the temporal consistency constraint is applied by carrying out single-view change detection independently on each original view ([2],[7]), thus computing a set of V binary change masks C_t^v, one for each original view $v = 1, \ldots, V$ (Figures 1(d-f)). The spatial coherence constraint is then applied by projecting all the change masks[1] into the virtual top-view, thus attaining a set of V top-view change masks $C_t^{v,T}$ (Figures 1(g-i)):

$$C_t^{v,T} = H^v(C_t^v) \tag{1}$$

Then, a common top-view change mask C_t^T is obtained by computing the intersection of all the top-view change masks (Figures 1(j)):

$$C_t^T = \bigcap_{v=1}^{V} C_t^{v,T} \tag{2}$$

The procedure outlined so far is substantially equivalent to the change maps fusion approach presented in [6]. The only difference is that we carry out change maps binarization directly as the final step of the temporal consistency constraint application. On the other hand, in [6] binarization is carried out in the virtual top-view after the spatial coherence constraint has been applied as well. We call "change masks fusion" this slightly different approach and "synergy mask" the binary mask of Equation 2. However, we deploy the *synergy* information within the top-view in a "dual" manner with respect to [6]. In fact, the synergy mask contains the pixels characterized by a high probability to be the image of a ground plane patch for which the emitted radiance is changed. These pixels correspond to people feet as well as to possible ground plane appearance changes, such as those due to shadows cast by people or to light spots hitting the ground plane. Therefore, instead of using the synergy mask to detect foreground objects ground locations (people feet), we use it to filter-out ground plane appearance changes, like shadows or light spots. In particular, instead of considering the synergy mask as the final output of the multi-view change detection, we back-project the synergy mask into all the original views, thus obtaining a

[1] actually, just the change masks portion inside the ground plane limits are projected

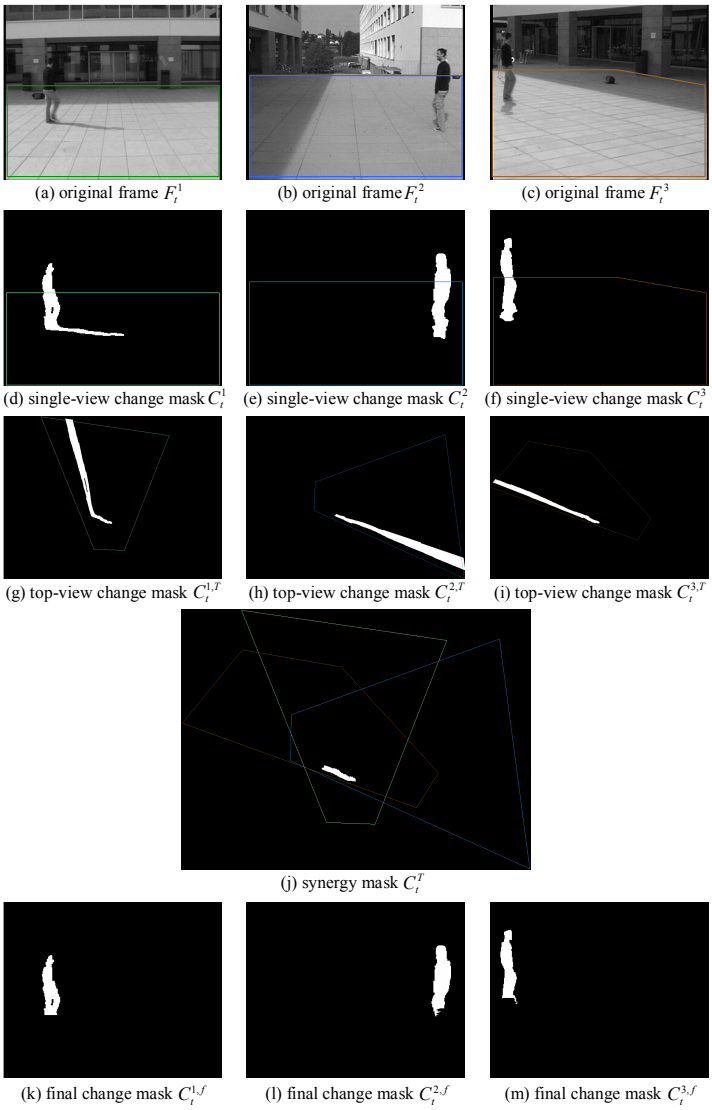


Figure 1: On-line main processing steps of the proposed multi-view approach.

set of V original view synergy masks $C_t^{T,v}$:

$$C_t^{T,v} = (H^v)^{-1}\left(C_t^T\right) \tag{3}$$

Then, for each view v we filter-out from the original view change mask C_t^v the foreground pixels belonging to the original view synergy mask $C_t^{T,v}$, thus attaining a set of V final change masks $C_t^{v,f}$ (Figures 1(k-m)):

$$C_t^{v,f}(\boldsymbol{p}^v) = \begin{cases} 0 & \text{if} \quad C_t^{T,v}(\boldsymbol{p}^v) = 1 \\ C_t^v(\boldsymbol{p}^v) & \text{otherwise} \end{cases} \tag{4}$$

Hence, another difference with respect to [6] is that we compute a set of V change masks, one for each original view, instead of a single change mask in the virtual top-view. Moreover, the change masks will include most of a person's body (ideally, the entire body but the feet). Unlike [5] and [8], our approach handles dynamic adjustments of camera parameters provided that a proper change detection algorithm (i.e. [2],[7]) is run on each original view. It is worth pointing out that algorithms such as [2] and [7] can also deal very effectively with sudden and global light changes.

The proposed approach is "general-purpose", in the sense that all the scene appearance changes detected by the employed single-view change detection algorithm which satisfy the spatial coherence constraint (i.e. which arise "near" the ground plane in a 3-dimensional sense) are filtered-out. In fact, no selectivity criterion is used in the removing rule of expression 4. In practice, just a geometrical constraint is applied, without considering any photometric information. On one hand this approach is general-purpose, but on the other hand a missed detections problem may arise due to the following two causes:

a) part of the foreground objects ground locations, especially people feet, may be removed together with the actual false changes (e.g. shadows) from the final change masks (Figure 1(k)). This is an inherent and easy to understand problem of the proposed approach, since ground locations of foreground objects yield appearance changes lying "near" the ground plane (i.e. they satisfy the spatial coherence constraint);

b) some "off-ground" portions of the foreground objects may be removed as well. This may occur for the original views in which the ground plane appearance changes are covered by foreground objects (Figure 1(l)). This is an inherent problem as well. In general, the higher the number of foreground objects present in the scene, the higher the probability of this problem to occur.

To face these two inherent problems we propose a less "general-purpose" removing rule, that we call "shadows-focused" removing rule. In fact, by this new rule we try to achieve a selective removal of just the ground plane appearance changes due to shadows. To this purpose, we exploit simple, well-known and commonly used photometric properties characterizing scene surfaces covered by shadows. The basic idea is that the measured intensity of a pixel imaging a scene background surface patch decreases according to a limited darkening factor d when covered by a cast shadow. Hence, the selective "shadows-focused" removing rule is the following:

$$C_t^{v,f}(\boldsymbol{p}) = \begin{cases} 0 & \text{if} \quad \left(C_t^{T,v}(\boldsymbol{p}^v) = 1\right) \wedge \left(d_{low} < \dfrac{F_t^v(\boldsymbol{p}^v)}{\hat{B}_t^v(\boldsymbol{p}^v)} < 1\right) \\ C_t^v(\boldsymbol{p}^v) & \text{otherwise} \end{cases} \tag{5}$$

where d_{low} is the lower darkening factor assumed for shadows effect and F_t^v, $\hat{B}_t^v$ are, respectively, the current frame and the current background model used by the single-view change detection algorithm in the original view v. In practice, for each view v the final change mask $C_t^{v,f}$ is not computed by filtering-out blindly all the foreground pixels of the original view synergy mask $C_t^{T,v}$ from the original view change mask C_t^v. Instead, just the foreground pixels satisfying the shadows photometric constraint are removed.

4 Experimental Results

Experiments have been carried out by running the proposed general-purpose and shadows-focused multi-view change detection approaches on several test video sequences. All the sequences have been captured by the same multi-view outdoor installation, consisting of three synchronized capturing devices imaging a common scene from very different view-points. Within the imaged scene, people walk and cast shadows on a planar ground. Here we present the change detection results for four different processing times (i.e. for four different triples of simultaneous frames) of a test sequence. In particular, the change masks computed by the general-purpose (blind removing rule of Equation 4) and by the shadows-focused (selective removing rule of Equation 5) approaches are directly compared in Figures 2-3. In particular, a value $d_{low} = 0.5$ is used in the shadows-focused removing rule. Shadows cast by moving people on the ground plane are removed effectively by both the approaches. In fact, since shadows seen in each view lie on the ground plane their entire shapes will be projected into the synergy mask and hence detected. This works well for long as well as short shadows. However, the general-purpose approach suffers from a missed detections problem, as expected. On one hand, in each view people feet may be partially removed, independently from the reciprocal position of people and cast shadows. In fact, feet yield a local change of the radiance emitted by the ground plane. As an example, the change masks on the left and on the right of the centre row of Figures 2(a,b) show how feet can be partially removed also in the very favourable situation of a single person moving in the scene without covering its cast shadow. On the other hand, "off-ground" portions of people's body may be removed as well when cast shadows are covered by people. This is the case of Figure 2(a), top row, in the middle, where the person covers almost completely its cast shadow. As a consequence, the lower portion of the person's body, that is the portion covering the cast shadow, is detected as unchanged, as shown in Figure 2(a), centre row, in the middle. In general, the higher the number of persons present in the scene, the higher the probability of this problem to occur, as shown in Figures 3(a,b), centre row. As for the considered test sequences, the missed detections problem is solved quite effectively by the shadows-focused approach, as regards both the feet and the covered shadows problems (Figures 2-3(a,b), bottom row). However, it is worth noticing that in general the persons' body appearance impacts the actual effectiveness of the shadows-focused approach in dealing with the missed detections problem. Finally, we point out that a shadow removal approach based only on the application of the photometric constraint in Equation 5 would be prone to the detection of false shadows not lying on the ground plane.

5 Conclusions

We have presented a multi-view change detection approach aimed at being robust to the major disturbance factors acting in real-world applications. On one hand, camera noise and disturbance factors yielding slow or global background appearance changes are dealt with by single-view change detection carried out independently on each original view. On the other hand, fast-and-local appearance changes are filtered-out by fusing the single-view change masks into a common virtual top-view and then back-projecting the attained synergy mask into the original views. However, sudden changes due to specular reflections can not be dealt with by the proposed algorithm for the ground plane constraint does

not hold in this case. Since a missed detections problem may arise due to causes which are inherent to the presented approach, a less general-purpose version of the algorithm has been proposed as well, focused on shadows removal. Since the appearance changes occurring in the available multi-view test sequences are all due to shadows cast by moving people on the ground plane, the shadows-focused approach yields better results than the general-purpose approach, as shown by experiments. Unlike other state-of-the-art multi-view change detection algorithms, which compute a single change mask in a reference ([5],[8]) or a virtual ([6]) view, the output of our approach is a set of different change masks, one for each original view. This output is suitable to be fed to a multi-view tracking algorithm such as ([1]).

References

[1] J. Berclaz, F. Fleuret, and P. Fua. Robust people tracking with global trajectory optimization. In *Proc. CVPR'06*, volume 1, pages 744–750, June 2006.

[2] A. Bevilacqua, L. Di Stefano, and A. Lanza. Coarse-to-fine strategy for robust and efficient change detectors. In *Proc. AVSS'05*, pages 87–92, September 2005.

[3] A. Elgammal, R. Duraiswami, D. Harwood, and L. S. Davis. Background and foreground modeling using nonparametric kernel density estimation for visual surveillance. *Proceedings of the IEEE*, 90(7):1151–1163, July 2002.

[4] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, second edition, 2004.

[5] Y. A. Ivanov, A. F. Bobick, and J. Liu. Fast lighting independent background subtraction. *International Journal of Computer Vision*, 37(2):199–207, June 2000.

[6] S. M. Khan and M. Shah. A multiview approach to tracking people in crowded scenes using a planar homography constraint. In *Proc. ECCV'06*, volume 4, pages 133–146, May 2006.

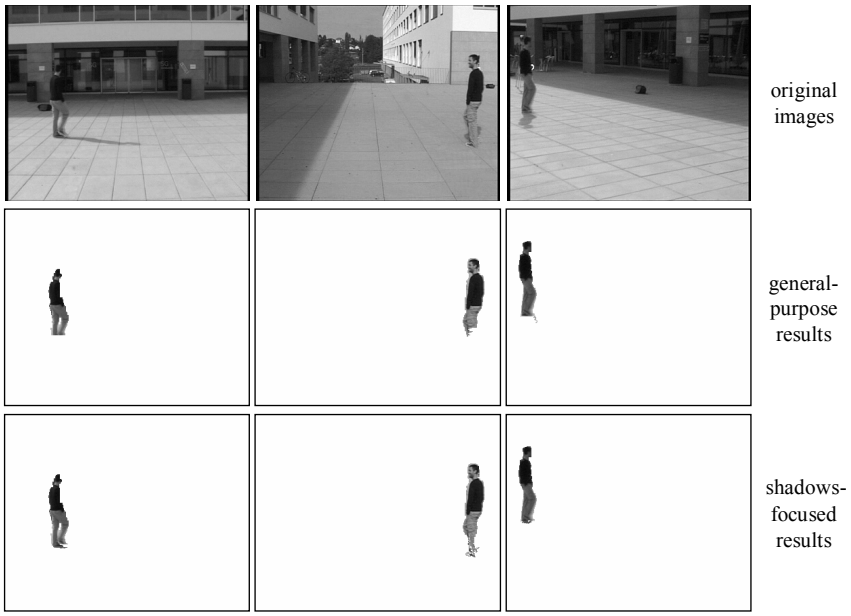
[7] A. Lanza and L. Di Stefano. Detecting changes in grey level sequences by ML isotonic regression. In *Proc. AVSS'06*, pages 4–4, November 2006.

[8] S. N. Lim, A. Mittal, L. S. Davis, and N. Paragios. Fast illumination-invariant background subtraction using two-views: Error analysis, sensor placement and applications. In *Proc. CVPR'05*, volume 1, pages 1071–1078, June 2005.

[9] N. Ohta. A statistical approach to background subtraction for surveillance systems. In *Proc. ICCV'01*, volume 2, pages 481–486, July 2001.

[10] C. Stauffer and W. E. L. Grimson. Adaptive background mixture models for real-time tracking. In *Proc. CVPR'99*, volume 2, pages 246–252, June 1999.

[11] B. Xie, V. Ramesh, and T. Boult. Sudden illumination change detection using order consistency. *Image and Vision Computing*, 22(2):117–125, February 2004.

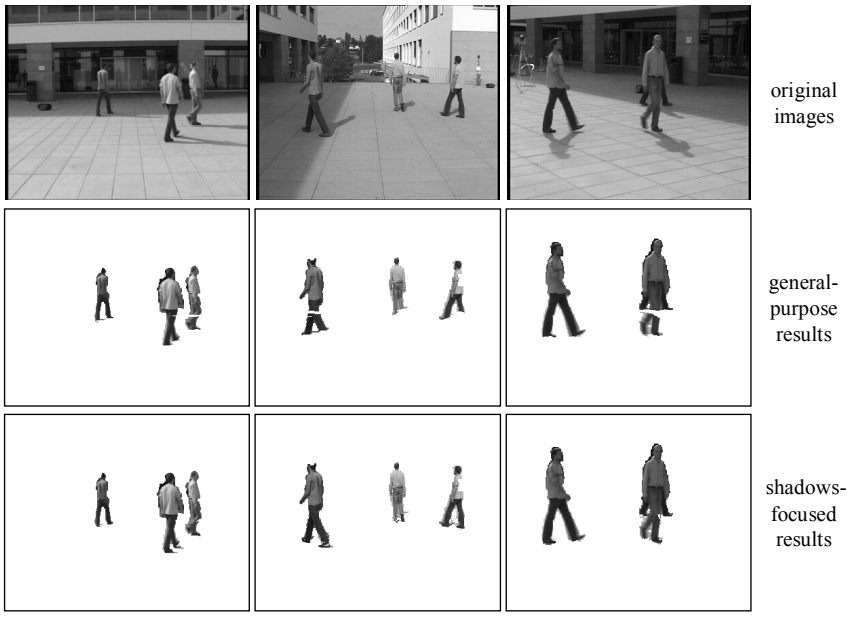


(a) frame 76

(b) frame 133

Figure 2: Change masks computed by the proposed general-purpose (centre row of (a) and (b)) and shadows-focused (bottom row of (a) and (b)) change detection approaches for frames 76 (top row of (a)) and 133 (top row of (b)).

(a) frame 333



(b) frame 355

Figure 3: Change masks computed by the proposed general-purpose (centre row of (a) and (b)) and shadows-focused (bottom row of (a) and (b)) change detection approaches for frames 333 (top row of (a)) and 355 (top row of (b)).